PATHCAST: PATHWAY-STRUCTURED GENE EXPRESSION FORECASTING IN CARDIAC ORGANOIDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Forecasting gene expression during cardiac organoid development can support experimental planning and the study of developmental programs. However, sparse, irregularly sampled RNA measurements provide limited evidence for the coordinated expression changes that emerge at later stages. We introduce PATHCAST, a pathway-structured Transformer that integrates biological pathway memberships and gene descriptions to forecast group-level expression beyond the observed RNA history. The central design couples complementary representations of pathway dynamics: a history branch uses temporal and cross-pathway differential attention, while a state branch summarizes reconstructable pathway states and recent trends. Their fusion predicts continuous-time state increments, which a pathway-constrained dynamic decoder translates into future expression changes with explicit pathway and gene-level contributions. Across within-condition and reference-assisted forecasting on three cardiac organoid datasets and one hepatic differentiation dataset, PATHCAST achieves the lowest overall mean squared error among the reported baselines. Ablations assess the contributions of pathway structure and complementary modeling components. Pathway analyses and uncertainty-based risk ranking characterize forecast behavior. Together, these findings support pathway-structured representations for interpretable forecasting from limited transcriptomic observations.[1]

1 INTRODUCTION

Human cardiac organoids provide a controllable system for investigating early heart development and congenital heart disease (Hofbauer et al., 2021), as well as drug responses in developing cardiac tissues (Volmert et al., 2023). Gene expression time courses connect molecular programs to cell differentiation and tissue formation (Lewis-Israeli et al., 2021). Yet repeated sampling and sequencing are costly, limiting the availability of dense transcriptomic time series (Alatkar & Wang, 2025). Forecasting future expression from a short observed history could therefore help prioritize follow-up measurements and experiments. This setting motivates models that exploit biological structure while supporting interpretation and assessment of predictive reliability.

The central modeling challenge is to extrapolate coordinated molecular changes from sparse, irregularly sampled observations. Thousands of genes evolve jointly, yet early expression histories may contain little variation even for genes that change substantially later. Biological pathways offer a structured representation of this coordination, but forecasting requires modeling how pathway representations evolve beyond the observation window. General forecasters such as iTransformer are evaluated on long, regularly sampled electricity, traffic, and weather series (Liu et al., 2024b). Cellular dynamics models such as ARTEMIS are also evaluated by reconstructing missing intermediate time points (Alatkar & Wang, 2025). Unlike interpolation, which can exploit observations on both sides of a missing time point, forecasting must predict beyond the available target-condition history.

A complementary challenge is to connect expression forecasts to biological structure without conflating interpretability with reliability. TOSICA uses pathway or regulon tokens for interpretable cell-type annotation (Chen et al., 2023), while SurvPath integrates pathway representations and histology for survival prediction (Jaume et al., 2024). These approaches explain annotation or prognostic outputs rather than contributions to future expression changes. Forecasting benefits from uncertainty

[1]We will release the code and data upon acceptance to support reproducibility.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

assessment to distinguish lower-risk predictions and expose failures across horizons (Gao et al., 2026). Crucially, pathway contributions describe model behavior rather than establish biological causation.

We introduce PATHCAST, a pathway-structured Transformer for forecasting group-level gene expression from limited RNA histories (Figure 1). Pathway memberships and gene descriptions inform two complementary branches: the history branch models temporal and cross-pathway dependencies with differential Transformers, while the state branch summarizes reconstructable pathway states and recent trends. Their fusion predicts continuous-time state increments, decoded into additive pathway, auxiliary, and gene-level expression contributions. Experiments cover within-condition and reference-assisted cardiac organoid forecasting, with reference profiles at forecast times available only in the latter, and an additional hepatic differentiation evaluation. The reported results show competitive accuracy, while ablations and uncertainty analyses assess model contributions and predictive reliability. Separate cross-study experiments demonstrate gains over training from scratch through pretraining and prediction-head adaptation of a direct-gene variant.

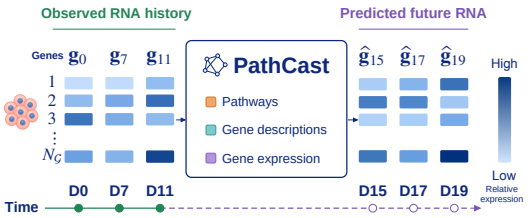

Figure 1: Overview of PATHCAST.

Our contributions are threefold: (1) We introduce a pathway-structured forecasting architecture that integrates biological knowledge with complementary history and state representations to predict future gene expression from sparse observations. (2) We formulate expression changes as explicit additive pathway, auxiliary, and gene-level contributions, enabling inspection of the model's forecasts. (3) We evaluate forecasting accuracy across cardiac organoid and hepatic differentiation datasets, with EB-hCO ablations and exploratory analyses of uncertainty-based risk ranking.

2 RELATED WORK

Pathway-Based Interpretability. Biological pathways provide structured representations that connect molecular measurements to recognizable biological processes and functions. TOSICA incorporates pathway structure into Transformers for interpretable cell-type annotation (Chen et al., 2023). For disease diagnosis and prognosis, Pathformer models interactions between pathways using multi-omics data (Liu et al., 2024a). SurvPath extends pathway-based modeling to multimodal analysis, learning interactions between pathway representations and histology for survival prediction (Jaume et al., 2024).

Language-Informed Gene Representations. Gene descriptions provide functional context that complements expression measurements. GenePT uses embeddings of gene descriptions to construct gene and cell representations (Chen & Zou, 2025). Language also supports cell-level analysis: LangCell aligns expression and text for cell identity understanding (Zhao et al., 2024), while CellWhisperer enables transcriptome exploration through natural language (Schaefer et al., 2025). Beyond representation learning, scGenePT incorporates gene-text embeddings into a perturbation model to predict cellular responses (Istrate et al., 2024).

Modeling Gene Expression Dynamics. Time-resolved transcriptomic data have motivated models that learn developmental dynamics from cell populations sampled at discrete time points. PRESCIENT learns stochastic differentiation landscapes to simulate cell trajectories (Yeo et al., 2021), while MIOFlow models continuous population flows along an expression manifold (Huguet et al., 2022). ARTEMIS combines autoencoders and Schrödinger bridges to recover expression and population changes at unmeasured time points (Alatkar & Wang, 2025). Alongside trajectory reconstruction, scMix explicitly evaluates future expression forecasting by integrating historical observations and irregular time intervals (Han & Kim, 2026).

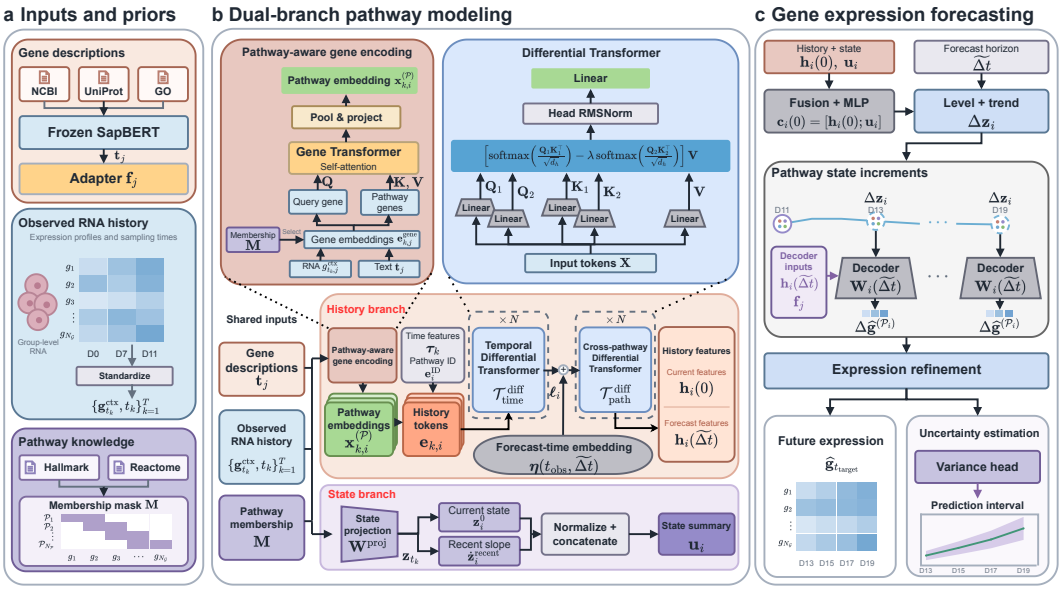

Figure 2: Architecture of PATHCAST. (a) Gene descriptions, observed RNA history and pathway memberships provide model inputs. (b) The history branch captures temporal and cross-pathway dependencies with differential Transformers, while the state branch summarizes pathway states and recent trends. (c) Fused features predict state increments for dynamic decoding into expression changes; a variance head estimates uncertainty.

3 METHOD

Problem formulation. Let $\mathbf{g}_{t_k} \in \mathbb{R}^{N_\mathcal{G}}$ denote a recorded group's profile at time t_k, with $N_\mathcal{G}$ genes. We observe T times $t_1 < \cdots < t_T = t_{\text{obs}}$ and predict $\widehat{\mathbf{g}}_{t_{\text{target}}}$ at $t_{\text{target}} > t_{\text{obs}}$. Adjacent intervals are $\Delta t_1 = 0$ and $\Delta t_k = t_k - t_{k-1}$; the forecast gap is $\Delta t_{\text{pred}} = t_{\text{target}} - t_{\text{obs}}$. We normalize time intervals by a fixed training-derived scale s_t, using $\widetilde{\Delta t} = \Delta t_{\text{pred}}/s_t$. Each target uses the same observed history, without feeding predictions back as observations. The indices i, j, and k refer to pathways or auxiliary components, genes, and observation times, respectively. Figure 2 provides an overview of the PATHCAST architecture.

3.1 KNOWLEDGE-INFORMED INPUT ENCODING

Gene descriptions. We combine NCBI Gene summaries, reviewed UniProt descriptions and Gene Ontology annotations (Istrate et al., 2024) for complementary information about genes, protein functions and biological context. A frozen SapBERT encoder $E_{\text{txt}}(\cdot)$ (Liu et al., 2021) embeds each available source; averaging yields the gene-text vector $\mathbf{t}_j$. A learned projection supplies $\mathbf{t}_j$ to the Gene Transformer, while a separate adapter combines text and expression statistics into d-dimensional gene features $\mathbf{f}_j$ for the reconstruction dictionary, dynamic decoder and gene-level correction. Sources and adapter details are in Appendices A and B.

Observed RNA history. The input history comprises profiles $\{\mathbf{g}^{\text{in}}_{t_k}\}_{k=1}^{T}$ at times t_k. Gene-wise means and scales s_j from permitted training profiles yield standardized inputs $\mathbf{g}^{\text{ctx}}_{t_k}$ shared by both branches.

Pathway knowledge. We use MSigDB Hallmark (Liberzon et al., 2015) and Reactome (Milacic et al., 2024) memberships to define a fixed matrix $\mathbf{M} \in \{0,1\}^{N_\mathcal{P} \times N_\mathcal{G}}$ for $N_\mathcal{P}$ retained pathways (Jaume et al., 2024), with $M_{ij} = 1$ exactly when $j \in \mathcal{P}_i$, the measured members of pathway i. This matrix restricts gene attention and the pathway-specific dictionary and decoder to member genes. Both branches additionally use N_{aux} auxiliary components to capture variation beyond annotated pathways (Seninge et al., 2021).

3.2 PATHWAY HISTORY MODELING

The history branch learns pathway representations from member genes and models their temporal evolution and interactions with differential Transformers.

Pathway-aware Gene Encoding. At time t_k, we combine observed gene expression with functional context from text to construct a gene token:

$$\mathbf{e}_{k,j}^{\text{gene}} = \text{LN}\big[\phi_{\text{expr}}\big(g_{t_k,j}^{\text{ctx}}\big) + \mathbf{W}_{\text{txt}}\mathbf{t}_j\big]\,. \tag{1}$$

ϕ_{expr} embeds gene j's standardized expression $g_{t_k,j}^{\text{ctx}}$, $\mathbf{W}_{\text{txt}}$ projects text, and LN denotes layer normalization.

Within each pathway $\mathcal{P}_i$, the shared Gene Transformer $\mathcal{T}_{\text{gene}}$ contextualizes member-gene tokens; mean pooling and projection yield a pathway representation:

$$\begin{aligned}\mathbf{H}_{k,i} &= \mathcal{T}_{\text{gene}}\Big([\mathbf{e}_{k,j}^{\text{gene}}]_{j\in\mathcal{P}_i}\Big)\,,\\ \mathbf{x}_{k,i}^{(\mathcal{P})} &= \frac{\mathbf{W}_{\text{pool}}}{|\mathcal{P}_i|}\sum_{j\in\mathcal{P}_i}\mathbf{H}_{k,i,j}.\end{aligned} \tag{2}$$

$\mathbf{H}_{k,i}$ stacks gene vectors $\mathbf{H}_{k,i,j}$, $|\mathcal{P}_i|$ counts genes, and $\mathbf{W}_{\text{pool}}$ is a learned projection.

History Token Construction. For temporal modeling, we construct one history token $\mathbf{e}_{k,i}$ per pathway or auxiliary component at each observed time. Each token combines a pathway embedding or dictionary-projected auxiliary state with its adjacent-time slope, running mean and sampling-time features, together with a learned component identity. Slopes and running means capture recent changes and historical levels using observations through t_k only; sampling-time features encode times and intervals.

Temporal and Cross-Pathway Differential Modeling. Strong temporal persistence and coordinated pathway activity may obscure dependencies predictive of expression increments. Both Transformers therefore use differential attention, subtracting a learned multiple of one attention map from another (Ye et al., 2025). This aims to suppress redundant shared attention patterns and highlight temporal dynamics beyond persistence and cross-pathway dependencies beyond broad co-variation. The shared temporal Differential Transformer $\mathcal{T}_{\text{time}}^{\text{diff}}$ summarizes each history; the cross-pathway Differential Transformer $\mathcal{T}_{\text{path}}^{\text{diff}}$ then processes these summaries with a forecast-time embedding:

$$\begin{aligned}\boldsymbol{\ell}_i &= \big[\mathcal{T}_{\text{time}}^{\text{diff}}([\mathbf{e}_{k,i}]_{k=1}^{T})\big]_T\,,\\ [\mathbf{h}_i(\widetilde{\Delta t})]_i &= \mathcal{T}_{\text{path}}^{\text{diff}}\Big([\boldsymbol{\ell}_i + \boldsymbol{\eta}(t_{\text{obs}}, \widetilde{\Delta t})]_i\Big)\,.\end{aligned} \tag{3}$$

Here, $[\cdot]_T$ selects the final observed output, giving the temporal summary $\boldsymbol{\ell}_i$, and $[\cdot]_i$ stacks component representations. The learned embedding $\boldsymbol{\eta}$ encodes the observation cutoff and normalized forecast horizon. The resulting history context $\mathbf{h}_i(\widetilde{\Delta t})$ incorporates forecast-time information and interactions across components. We use the zero-horizon context $\mathbf{h}_i(0)$ for fusion with current-state information and $\mathbf{h}_i(\widetilde{\Delta t})$ to condition the dynamic decoder at the requested horizon. Token construction and attention implementation are detailed in Appendix B.

3.3 PATHWAY STATE MODELING

The state branch projects observed expression into pathway states and summarizes their current values and recent changes to condition the prediction of state changes.

Reconstructable pathway states. We represent each annotated pathway by a K-dimensional latent state and each auxiliary component by a scalar state, giving $N_z = KN_{\mathcal{P}} + N_{\text{aux}}$ latent dimensions. A reconstruction dictionary $\mathbf{W}^{\text{base}} \in \mathbb{R}^{N_z\times N_{\mathcal{G}}}$ has one row per state coordinate and one column per gene; its transpose maps states back to standardized expression. Each pathway occupies K rows with connections restricted to its member genes, while auxiliary rows connect to all genes. Gene features $\mathbf{f}_j$ adjust the pathway weights, with dictionary parameters learned separately from the

Gene Transformer. We infer the full state vector $\mathbf{z}_{t_k} \in \mathbb{R}^{N_z}$ from the expression context through a joint regularized projection:

$$\begin{aligned} \mathbf{z}_{t_k} &= \mathbf{W}^{\mathrm{proj}} \mathbf{g}^{\mathrm{ctx}}_{t_k}, \\ \mathbf{W}^{\mathrm{proj}} &= \left[\mathbf{W}^{\mathrm{base}} (\mathbf{W}^{\mathrm{base}})^\top + \lambda_{\mathrm{proj}} \mathbf{\Lambda}\right]^{-1} \mathbf{W}^{\mathrm{base}}. \end{aligned} \tag{4}$$

Here, $\mathbf{W}^{\mathrm{proj}} \in \mathbb{R}^{N_z \times N_{\mathcal{G}}}$ is the projection derived from the dictionary, $\lambda_{\mathrm{proj}} > 0$ controls regularization, and $\mathbf{\Lambda}$ is a diagonal scaling matrix. Jointly inferring all states accounts for overlap among dictionary rows. The K-dimensional state vector for pathway i is $\mathbf{z}_{t_k,i}$. Joint projection can make these states depend on nonmember genes; they do not directly measure biochemical activity.

State features. For pathway i, the current state $\mathbf{z}_i^0 = \mathbf{z}_{t_{\mathrm{obs}},i}$ is inferred from the latest expression context. Its recent slope $\dot{\mathbf{z}}_i^{\mathrm{recent}}$ is the difference between the last two observed states divided by their normalized sampling gap. Normalizing and concatenating these two vectors gives the state summary $\mathbf{u}_i \in \mathbb{R}^{2K}$, which conditions the prediction heads. Dictionary construction and state-feature normalization are detailed in Appendix B.

3.4 History–State Fusion and Forecasting

History–State Fusion. We concatenate the history context and state summary as $\mathbf{c}_i(0) = [\mathbf{h}_i(0); \mathbf{u}_i]$. MLP heads map this fused representation to pathway trajectory parameters, which are shared across forecast horizons for the same observed history.

Level–Trend Forecasting. Each pathway-state coordinate $a = 1, \ldots, K$ combines a saturating level adjustment and a sustained trend, using $\kappa(\widetilde{\Delta t}; \tau) = 1 - \exp(-\widetilde{\Delta t}/\tau)$, with time constant $\tau > 0$:

$$\begin{aligned} \Delta z_{ia}^{\mathrm{level}} &= \beta_{ia} \kappa(\widetilde{\Delta t}; \tau_{ia}^{\mathrm{level}}), \\ \Delta z_{ia}^{\mathrm{trend}} &= \alpha_i \left[v_{ia} \widetilde{\Delta t} + \delta v_{ia} \tau_{ia}^{\mathrm{trend}} \kappa(\widetilde{\Delta t}; \tau_{ia}^{\mathrm{trend}}) \right]. \end{aligned} \tag{5}$$

β_{ia} is the signed level amplitude, v_{ia} the sustained velocity, and δv_{ia} its initial offset; $\tau_{ia}^{\mathrm{level}}$ and $\tau_{ia}^{\mathrm{trend}}$ set the level and velocity timescales, and $\alpha_i \in (0, 1)$ gates the trend.

The total increment $\Delta \mathbf{z}_i = \Delta \mathbf{z}_i^{\mathrm{level}} + \Delta \mathbf{z}_i^{\mathrm{trend}}$ is passed to the dynamic decoder in Section 3.5. Auxiliary components use only the level adjustment.

3.5 Gene Expression Prediction and Uncertainty

We decode predicted state increments into gene expression forecasts and quantify their uncertainty.

Dynamic pathway decoding. History context $\mathbf{h}_i(\widetilde{\Delta t})$ and gene features $\mathbf{f}_j$ adapt pathway i's dictionary block $\mathbf{W}_i^{\mathrm{base}}$ into decoder weights $\mathbf{W}_i(\widetilde{\Delta t})$, restricted to member genes. These map state increments $\Delta \mathbf{z}_i$ to expression contributions:

$$\Delta \widehat{\mathbf{g}}^{(\mathcal{P}_i)} = \mathbf{s} \odot \left[\mathbf{W}_i(\widetilde{\Delta t})^\top \Delta \mathbf{z}_i \right]. \tag{6}$$

Here, $\mathbf{s} = [s_j]_j$ contains the gene-wise training scales and $\odot$ denotes element-wise multiplication, restoring increments to the input-expression scale. The dense auxiliary dictionary block similarly decodes auxiliary-state changes into $\Delta \widehat{\mathbf{g}}^{\mathrm{aux}}$.

Gene-level correction and prediction. A shared MLP uses gene features and observed-history information to produce a signed, saturating correction for each gene. These corrections $\Delta \widehat{\mathbf{g}}^{\mathrm{gene}}$ and the auxiliary outputs supplement the named pathway contributions. Adding all components to the task-specific base profile $\mathbf{g}^{\mathrm{base}}$ gives the forecast:

$$\widehat{\mathbf{g}}_{t_{\mathrm{target}}} = \mathbf{g}^{\mathrm{base}} + \sum_{i=1}^{N_{\mathcal{P}}} \Delta \widehat{\mathbf{g}}^{(\mathcal{P}_i)} + \Delta \widehat{\mathbf{g}}^{\mathrm{aux}} + \Delta \widehat{\mathbf{g}}^{\mathrm{gene}}. \tag{7}$$

All increments vanish at zero horizon. Decoder and correction details are given in Appendix B.

Forecasting objective. With branch inputs restricted to the observed prefix, we jointly optimize the forecasting modules and reconstruction dictionary to predict later training targets:

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{dyn}} + \lambda_{\text{recon}}\mathcal{L}_{\text{recon}} + \mathcal{R}. \tag{8}$$

$\mathcal{L}_{\text{pred}}$ penalizes expression errors; $\mathcal{L}_{\text{dyn}}$ supervises adjacent target rates and first–last changes for a shared history. $\mathcal{L}_{\text{recon}}$ reconstructs observations with weight λ_{recon}, and $\mathcal{R}$ regularizes output amplitudes and decoder adjustments. Details are in Appendix C.

Uncertainty Estimation and Calibration. For a single forecasting model, a shared variance head estimates the residual variance $\sigma_j^2(t)$ for gene j at target time t. Its Gaussian negative log-likelihood objective updates only the variance head (Stirn et al., 2023). We calibrate prediction intervals using historical residuals (Kuleshov et al., 2018). For nominal coverage p, the interval is

$$I_{p,j}(t) = [\widehat{g}_{t,j} - q_p\sigma_j(t),\ \widehat{g}_{t,j} + q_p\sigma_j(t)]\,. \tag{9}$$

Here, $\widehat{g}_{t,j}$ and $\sigma_j(t)$ denote the point forecast and predicted residual standard deviation; q_p is the empirical p-quantile of absolute historical residuals divided by their predicted standard deviations. Head training, calibration and the three-seed ensemble used in the reported uncertainty experiments are detailed in Appendix D. Historical calibration does not guarantee coverage at longer horizons.

4 Experiments

4.1 Experimental setup

Datasets. We evaluate four public bulk RNA-seq datasets: three cardiac organoid time courses and one hepatic differentiation series. We refer to them as **EB-hCO** (embryoid-body-derived human cardiac organoids; GEO: GSE153185) (Lewis-Israeli et al., 2021); **Cardioids** (self-organizing cardiac organoids; GEO: GSE148025) (Hofbauer et al., 2021); **cVO** (cardiac vascularized organoids; GEO: GSE185194) (Abilez et al., 2025); and **H1-Hep** (hepatic differentiation of H1 human embryonic stem cells; GEO: GSE70741) (Li et al., 2017). Cardioids includes CHIR4 and CHIR8 (4 and 8 μM CHIR99021); cVO includes standard-differentiation controls (Unvasc.) and vascular-induction samples (Vasc.). All forecasts target normalized expression over 2,000 genes. H1-Hep provides a non-cardiac evaluation (Section 4.2.4).

Baselines. We compare four types of baselines: simple extrapolation, using last-velocity; general time-series forecasters, including iTransformer (Liu et al., 2024b), ReIMTS–GraFITi (Li et al., 2026; Yalavarthi et al., 2024) and ASTGI (Liu et al., 2026); pretrained foundation models, including scGPT-temporal (Cui et al., 2024) and Time-MoE-50M (Shi et al., 2025); and gene-expression models adapted for temporal profile forecasting, including MIOFlow (Huguet et al., 2022), ARTEMIS (Alatkar & Wang, 2025), and scMix (Han & Kim, 2026).

Metrics. We report mean squared error (MSE), mean absolute error (MAE), Pearson correlation across genes (r), and Top-50 MSE ($\text{MSE}_{\text{Top50}}$). We select the 50 genes with the highest variance of expression changes between consecutive observed time points and compute Top-50 MSE over these genes. Lower MSE, MAE, and Top-50 MSE and higher r indicate better predictions.

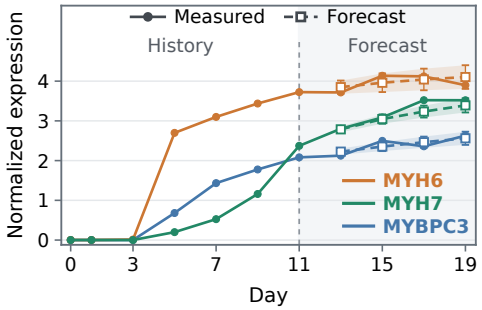

Figure 3: Observed and predicted gene expression trajectories on EB-hCO.

4.2 Forecasting performance

4.2.1 Within-condition forecasting

We evaluate within-condition forecasting on EB-hCO and Cardioids (CHIR4 and CHIR8), without target-day expression as input. Observation and forecast windows are provided in Table E1 in Appendix E. As shown in Table 1, Ours performs well across most evaluation metrics in all three settings. To examine gene-level trajectories,

we select MYH6, MYH7, and MYBPC3 as representative cardiac contractile genes (Abilez et al., 2025). MYH6 and MYH7 are used to characterize cardiomyocyte phenotypes in developing cardiac organoids, while MYBPC3 encodes a sarcomeric protein that regulates cardiac contraction. Figure 3 shows that the EB-hCO forecasts track the measured trajectories, capturing the overall expression trends and relative levels of the three genes.

Table 1: Within-condition forecasting on EB-hCO and Cardioids. Bold and underlined values indicate the best and second-best results.

	EB-hCO				Cardioids (CHIR4)				Cardioids (CHIR8)			
Method	MSE ↓	MAE ↓	MSE_{Top50} ↓	r ↑	MSE ↓	MAE ↓	MSE_{Top50} ↓	r ↑	MSE ↓	MAE ↓	MSE_{Top50} ↓	r ↑
Last-velocity	0.1126	0.1896	0.2628	0.9250	0.2585	0.2501	1.7641	0.8624	0.4107	0.3187	2.0640	0.8587
iTransformer	0.0502	0.1425	<u>0.1175</u>	0.9494	<u>0.0346</u>	<u>0.1004</u>	<u>0.1559</u>	<u>0.9591</u>	<u>0.0503</u>	<u>0.1304</u>	<u>0.1508</u>	<u>0.9440</u>
ReIMTS–GraFITi	0.1834	0.2688	1.4954	0.8203	0.0582	0.1438	0.4518	0.9323	0.1274	0.2128	1.0316	0.8938
ASTGI	0.1236	0.2237	0.7039	0.8715	0.1117	0.1874	0.7207	0.8788	0.1278	0.2137	0.9310	0.8690
scGPT-temporal	<u>0.0407</u>	<u>0.1333</u>	**0.1012**	<u>0.9649</u>	0.0691	0.1511	0.2338	0.9493	0.1504	0.2142	0.6146	0.9323
Time-MoE-50M	0.0634	0.1616	0.2947	0.9337	0.0654	0.1356	0.3474	0.9216	0.0970	0.1808	0.4682	0.8864
MIOFlow	0.7548	0.5245	6.2502	0.6820	0.7554	0.4514	4.7523	0.7497	1.7873	0.5804	12.8154	0.7924
ARTEMIS	0.1337	0.2484	0.4351	0.8610	0.0983	0.1834	0.6342	0.8825	0.1060	0.2003	0.5023	0.8814
scMix	0.1565	0.2800	0.4448	0.8453	0.1262	0.2257	0.9345	0.8501	0.2013	0.2883	1.3540	0.7704
Ours	**0.0322**	**0.1117**	0.1275	**0.9680**	**0.0282**	**0.0913**	**0.0988**	**0.9668**	**0.0425**	**0.1185**	**0.0961**	**0.9549**

4.2.2 Reference-assisted forecasting

Building on same-condition extrapolation, we ask whether measurements from a related condition can improve forecasting when the target condition is observed only over an early time window. We consider paired-condition time courses with reference measurements available at the forecast times. We evaluate both directions between Unvasc. and Vasc. in cVO and between CHIR4 and CHIR8 in Cardioids. In Table 2, Ours achieves the lowest MSE across all four forecasting directions. As shown in Figure 4, adding reference measurements reduces MSE by 25.58% for cVO Unvasc.→Vasc. relative to target-history-only forecasting. For Cardioids CHIR4→CHIR8, reference measurements reduce MSE by 21.47% relative to the same baseline.

Table 2: Reference-assisted forecasting on cVO and Cardioids.

	cVO								**Cardioids**							
	Unvasc. → Vasc.				Vasc. → Unvasc.				CHIR4 → CHIR8				CHIR8 → CHIR4			
Method	MSE ↓	MAE ↓	MSE_{Top50} ↓	r ↑	MSE ↓	MAE ↓	MSE_{Top50} ↓	r ↑	MSE ↓	MAE ↓	MSE_{Top50} ↓	r ↑	MSE ↓	MAE ↓	MSE_{Top50} ↓	r ↑
Last-velocity	0.1561	0.2057	1.5151	0.8998	0.1596	0.2075	1.5673	0.8776	0.3322	0.3023	1.7327	0.8489	0.3267	0.2999	1.6665	0.8019
ReIMTS–GraFITi	0.0604	0.1388	0.3429	0.9412	0.0644	0.1455	0.3876	0.9349	0.0556	0.1460	0.2809	0.9447	0.0646	0.1605	0.3441	0.9227
ASTGI	<u>0.0517</u>	<u>0.1266</u>	<u>0.2534</u>	<u>0.9506</u>	<u>0.0549</u>	<u>0.1363</u>	<u>0.2552</u>	0.8550	0.0453	0.1294	0.2224	<u>0.9529</u>	0.0458	0.1322	0.2357	0.9447
scGPT-temporal	0.1401	0.2341	1.0173	0.8994	0.1154	0.2072	0.7789	0.8980	0.1088	0.1875	0.3539	0.9337	0.0980	0.1868	0.3183	0.8935
Time-MoE-50M	0.0659	0.1411	0.4517	0.9347	0.0628	0.1395	0.3662	<u>0.9366</u>	<u>0.0429</u>	<u>0.1268</u>	<u>0.1822</u>	0.9524	<u>0.0446</u>	<u>0.1254</u>	<u>0.2236</u>	<u>0.9464</u>
MIOFlow	3.6586	0.6164	34.6320	0.8068	0.5143	0.2951	5.7322	0.8198	0.1609	0.2353	1.0754	0.8759	0.4095	0.3263	3.5496	0.7863
ARTEMIS	0.0708	0.1507	0.4647	0.9305	0.0681	0.1502	0.3768	0.9312	0.0630	0.1531	0.3752	0.9314	0.0562	0.1464	0.2596	0.9322
scMix	0.0771	0.1671	0.5142	0.9238	0.0760	0.1674	0.4891	0.9232	0.0814	0.1907	0.4135	0.9110	0.0822	0.1920	0.4057	0.9035
Ours	**0.0509**	**0.1258**	**0.2389**	**0.9520**	**0.0505**	**0.1249**	**0.2376**	**0.9497**	**0.0338**	**0.1081**	**0.1014**	**0.9667**	**0.0335**	**0.1076**	**0.1048**	**0.9599**

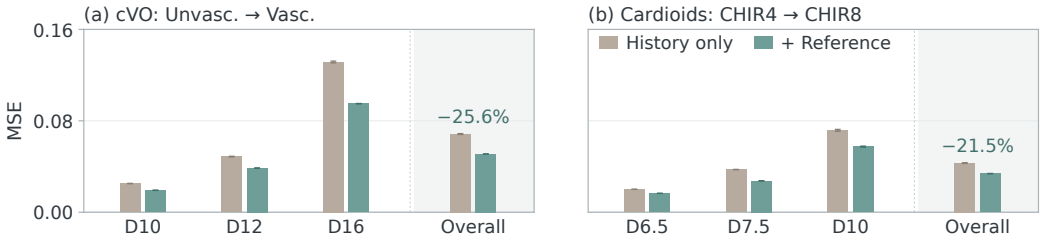

Figure 4: Effect of reference measurements on forecasting accuracy.

4.2.3 CROSS-STUDY TRANSFER

We evaluate whether cross-study pretraining improves forecasting with limited early observations. We pretrain the direct-gene variant on Cardioids and cVO and adapt it to EB-hCO using either the full D0–D11 history or a reduced D0–D7 history, while predicting D13–D19. As shown in Table 3, head adaptation outperforms training from scratch in both settings, demonstrating that cross-study knowledge improves forecasting when early observations are limited.

Table 3: Cross-study adaptation to EB-hCO.

	D11 history				D7 history			
Strategy	MSE ↓	MAE ↓	$\text{MSE}_{\text{Top50}}$ ↓	r ↑	MSE ↓	MAE ↓	$\text{MSE}_{\text{Top50}}$ ↓	r ↑
From scratch	0.0528	0.1528	0.2647	0.9449	0.1571	0.2818	0.4250	0.8253
Frozen transfer	0.0383	0.1312	0.0832	0.9588	0.0952	0.2136	0.2269	0.8921
Full adaptation	0.0361	0.1277	0.0809	0.9632	0.0925	0.2166	0.2226	**0.8945**
Head adaptation	**0.0310**	**0.1158**	**0.0676**	**0.9660**	**0.0913**	**0.2134**	**0.2179**	0.8943

4.2.4 FORECASTING DURING HEPATIC DIFFERENTIATION

To assess whether our model performs well beyond cardiac organoids, we evaluate forecasting on H1-Hep, a hepatic differentiation dataset. We select genes and estimate expression statistics from observed profiles, then evaluate forecasts against the later measurements. In Table 4, Ours achieves the best values on all four metrics among the listed methods, reducing MSE by 12.47% relative to iTransformer. This exploratory result supports applicability to hepatic differentiation.

Table 4: Forecasting performance on H1-Hep.

Method	MSE ↓	MAE ↓	$\text{MSE}_{\text{Top50}}$ ↓	r ↑
Last-velocity	0.1564	0.2382	1.2266	0.9069
iTransformer	0.0535	0.1568	0.3349	0.9472
ReIMTS–GraFITi	0.1014	0.2209	0.6396	0.9187
ASTGI	0.1099	0.2308	0.7122	0.8906
scGPT-temporal	0.1058	0.2187	0.7337	0.9258
Time-MoE-50M	0.1468	0.2451	1.5175	0.8391
MIOFlow	2.0481	0.8649	19.7956	0.7173
ARTEMIS	0.1905	0.3049	1.4022	0.7872
scMix	0.2553	0.3511	2.0414	0.7119
Ours	**0.0469**	**0.1471**	**0.2659**	**0.9529**

4.3 INTERPRETABILITY THROUGH PATHWAY CONTRIBUTIONS

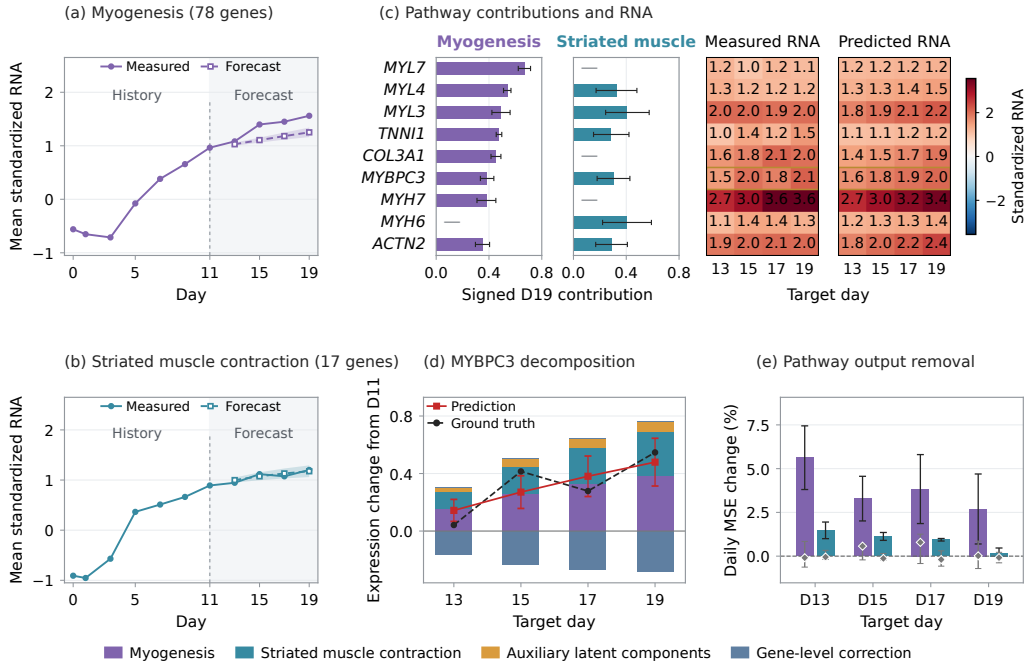

Figure 5: Pathway contributions to EB-hCO forecasts. (a–b) Mean expression of pathway genes. (c) Pathway contributions and gene expression. (d) MYBPC3 forecast decomposition. (e) MSE changes after pathway-output removal.

Figure 5a–b shows that forecasts capture the increasing mean expression of genes in the Myogenesis and Striated muscle contraction pathways in EB-hCO. Figure 5c highlights large contributions to MYL7 from Myogenesis and to MYL3 and MYH6 from Striated muscle contraction. For MYBPC3, Figure 5d shows that positive pathway contributions are partially offset by gene-level corrections. Removing the two pathway outputs increases MSE by 3.31% and 0.69%, respectively, as shown in Figure 5e, supporting their contribution to forecasting accuracy.

4.4 Uncertainty analysis

We evaluate gene-level uncertainty on EB-hCO to assess whether it identifies more reliable forecasts. Figure 6a shows that higher predicted uncertainty is associated with larger prediction errors. In Figure 6b, the 10% of genes with the lowest predicted standard deviation have an MSE approximately 92–95% lower than that across all genes.

4.5 Ablation studies

We examine the contribution of individual components on EB-hCO, as summarized in Table 5. ❶ replaces pathway modeling with direct-gene modeling, leading to higher errors and lower correlation. This supports the use of pathway representations for expression forecasting. ❷ removes state conditioning. The modest decline in accuracy suggests that explicit state information complements the history representation. ❸ removes history conditioning. This causes the largest performance drop, with MSE rising from 0.0322 to 0.0924, highlighting the importance of temporal context for forecasting future expression. ❹ removes the sustained trend while retaining level changes. The smaller decline across all metrics indicates that sustained trends provide additional predictive value. ❺ removes the gene-level correction. Accuracy also declines modestly, suggesting that gene-specific adjustments complement the shared pathway predictions. ❻ removes auxiliary forecast outputs. The marked MSE increase to 0.0472 suggests that auxiliary components contribute predictive information beyond the annotated pathways. ❼ removes all text modules, resulting in a modest decline in accuracy relative to its text-enabled baseline.

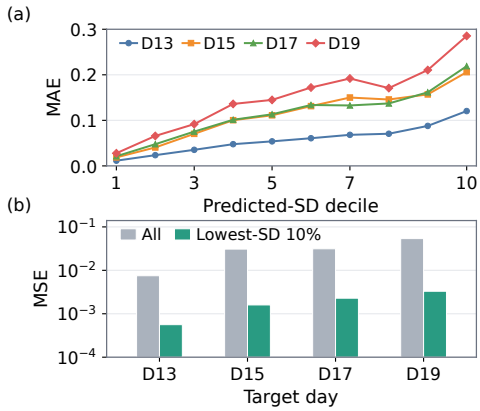

Figure 6: Uncertainty on EB-hCO: (a) risk ranking; (b) selective prediction.

Table 5: Ablation study on EB-hCO with D11 history and D13–D19 forecasts. Bold indicates the best result among Ours and settings ❶–❼.

Setting	MSE ↓	MAE ↓	$\text{MSE}_{\text{Top50}}$ ↓	r ↑
❶	0.0444	0.1340	0.2194	0.9551
❷	0.0337	0.1141	0.1420	0.9671
❸	0.0924	0.2094	0.3260	0.9020
❹	0.0341	0.1168	0.1522	0.9660
❺	0.0339	0.1158	0.1510	0.9665
❻	0.0472	0.1471	0.1350	0.9514
❼	0.0333	0.1142	**0.1207**	0.9674
Ours	**0.0322**	**0.1117**	0.1275	**0.9680**

5 Conclusion

We presented PathCast, a pathway-structured Transformer that combines history and state representations to forecast gene expression from limited RNA observations. Its additive pathway, auxiliary, and gene-level contributions make forecasts easier to interpret. Across cardiac organoid and hepatic differentiation datasets, PathCast achieves the lowest overall MSE among the reported baselines. Uncertainty-based risk ranking helps identify lower-error predictions. Future work will focus on independent validation, forecasting non-monotonic expression changes, and improving uncertainty calibration.

AI USE STATEMENT

Generative AI was used to help improve the wording and readability of the manuscript. The authors reviewed all AI-assisted edits and are responsible for the final version of the paper.

ETHICS STATEMENT

This study reanalyzes public bulk RNA-seq datasets from cardiac organoid and hepatic differentiation experiments. No new human-participant or animal experiments were conducted for this work. Predictions concern group-level expression in these experimental systems. Pathway contributions and component-removal analyses describe model behavior and do not establish biological causality or clinical validity.

REPRODUCIBILITY STATEMENT

Dataset accession numbers and evaluation settings are provided in Section 4.1. Appendix A documents the gene-description sources. Appendix B details preprocessing, model architecture, optimization settings, and random seeds. Appendix C specifies the training objective and regularization, while Appendix D describes uncertainty estimation and interval calibration. Observation and forecast windows and cross-study adaptation settings are provided in Appendix E.

REFERENCES

Oscar J. Abilez, Huaxiao Yang, Yuan Guan, Mengcheng Shen, Zehra Yildirim, Yan Zhuge, Ravichandra Venkateshappa, Shane R. Zhao, Angello H. Gomez, Marcel El-Mokahal, Logan Dunkenberger, Yoshikazu Ono, Masafumi Shibata, Peter N. Nwokoye, Lei Tian, Kitchener D. Wilson, Evan H. Lyall, Fangjun Jia, Hung Ta Wo, Gao Zhou, Bryan Aldana, Ioannis Karakikes, Detlef Obal, Gary Peltz, Christopher K. Zarins, and Joseph C. Wu. Gastruloids enable modeling of the earliest stages of human cardiac and hepatic vascularization. *Science*, 388(6751):eadu9375, 2025. doi: 10.1126/science.adu9375.

Sayali Anil Alatkar and Daifeng Wang. ARTEMIS integrates autoencoders and Schrödinger bridges to predict continuous dynamics of gene expression, cell population, and perturbation from time-series single-cell data. *Bioinformatics*, 41(Supplement 1):i189–i197, 2025. doi: 10.1093/bioinformatics/btaf218.

Jiawei Chen, Hao Xu, Wanyu Tao, Zhaoxiong Chen, Yuxuan Zhao, and Jing-Dong J. Han. Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1):223, 2023. doi: 10.1038/s41467-023-35923-4.

Yiqun Chen and James Zou. Simple and effective embedding model for single-cell biology built from ChatGPT. *Nature Biomedical Engineering*, 9(4):483–493, 2025. doi: 10.1038/s41551-024-01284-6.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

Dawei Gao, Angello Huerta Gomez, Mingchen Li, Marcel El-Mokahal, Huaxiao Yang, and Yunhe Feng. OrgaCast: A trustworthy spatiotemporal diffusion model for fluorescence organoid forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(6):4104–4112, 2026. doi: 10.1609/aaai.v40i6.42405. URL https://ojs.aaai.org/index.php/AAAI/article/view/42405.

Shangjin Han and Dongsup Kim. scMix: learning temporal dynamics of gene expression under irregular time intervals. *Bioinformatics*, 42(3):btag080, 2026. doi: 10.1093/bioinformatics/btag080.

Pablo Hofbauer, Stefan M. Jahnel, Nora Papai, Magdalena Giesshammer, Alison Deyett, Clara Schmidt, Mirjam Penc, Katherina Tavernini, Nastasja Grdseloff, Christy Meledeth, Lavinia Ceci Ginistrelli, Claudia Ctortecka, Šejla Šalic, Maria Novatchkova, and Sasha Mendjan. Cardioids reveal self-organizing principles of human cardiogenesis. *Cell*, 184(12):3299–3317.e22, 2021. doi: 10.1016/j.cell.2021.04.034.

Guillaume Huguet, Daniel Sumner Magruder, Alexander Tong, Oluwadamilola Fasina, Manik Kuchroo, Guy Wolf, and Smita Krishnaswamy. Manifold interpolating optimal-transport flows for trajectory inference. In *Advances in Neural Information Processing Systems*, volume 35, pp. 29705–29718. Curran Associates, Inc., 2022. doi: 10.52202/068431-2154.

Ana-Maria Istrate, Donghui Li, and Theofanis Karaletsos. scGenePT: Is language all you need for modeling single-cell perturbations? *bioRxiv*, 2024. doi: 10.1101/2024.10.23.619972. URL https://doi.org/10.1101/2024.10.23.619972. Preprint.

Guillaume Jaume, Anurag Vaidya, Richard J. Chen, Drew F. K. Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11579–11590, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Jaume_Modeling_Dense_Multimodal_Interactions_Between_Biological_Pathways_and_Histology_for_CVPR_2024_paper.html.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804. PMLR, 2018. URL https://proceedings.mlr.press/v80/kuleshov18a.html.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Yonatan R. Lewis-Israeli, Aaron H. Wasserman, Mitchell A. Gabalski, Brett D. Volmert, Yixuan Ming, Kristen A. Ball, Weiyang Yang, Jinyun Zou, Guangming Ni, Natalia Pajares, Xanthippi Chatzistavrou, Wen Li, Chao Zhou, and Aitor Aguirre. Self-assembling human heart organoids for the modeling of cardiac development and congenital heart disease. *Nature Communications*, 12(1):5142, 2021. doi: 10.1038/s41467-021-25329-5.

Boyuan Li, Zhen Liu, Yicheng Luo, and Qianli Ma. Learning recursive multi-scale representations for irregular multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 38480–38504, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/410dacbcfda0388f69b8983ed771c44e-Paper-Conference.pdf.

Qiuhong Li, Andrew P. Hutchins, Yong Chen, Shengbiao Li, Yongli Shan, Baojian Liao, Dejin Zheng, Xi Shi, Yinxiong Li, Wai-Yee Chan, Guangjin Pan, Shicheng Wei, Xiaodong Shu, and Duanqing Pei. A sequential EMT-MET mechanism drives the differentiation of human embryonic stem cells towards hepatocytes. *Nature Communications*, 8(1):15166, 2017. doi: 10.1038/ncomms15166. URL https://www.nature.com/articles/ncomms15166.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6): 417–425, 2015. doi: 10.1016/j.cels.2015.12.004.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4228–4238, 2021. doi: 10.18653/v1/2021.naacl-main.334. URL https://aclanthology.org/2021.naacl-main.334/.

Xiaofan Liu, Yuhuan Tao, Zilin Cai, Pengfei Bao, Hongli Ma, Kexing Li, Mengtao Li, Yunping Zhu, and Zhi John Lu. Pathformer: a biological pathway informed transformer for disease diagnosis and prognosis using multi-omics data. *Bioinformatics*, 40(5):btae316, 2024a. doi: 10.1093/bioinformatics/btae316.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Chenjuan Guo, Bin Yang, and Jilin Hu. ASTGI: Adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. In *International Conference on Learning Representations*, volume 2026, pp. 23940–23958, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/28d38c036365420f61ce03300418e44a-Paper-Conference.pdf.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, volume 2024, pp. 11116–11140, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/2ea18fdc667e0ef2ad82b2b4d65147ad-Paper-Conference.pdf.

Marija Milacic, Deidre Beavers, Patrick Conley, Chuqiao Gong, Marc Gillespie, Johannes Griss, Robin Haw, Bijay Jassal, Lisa Matthews, Bruce May, Robert Petryszak, Eliot Ragueneau, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Ralf Stephan, Krishna Tiwari, Thawfeek Varusai, Joel Weiser, Adam Wright, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The Reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52(D1): D672–D678, 2024. doi: 10.1093/nar/gkad1025.

Moritz Schaefer, Peter Peneder, Daniel Malzl, Salvo Danilo Lombardo, Mihaela Peycheva, Jake Burton, Anna Hakobyan, Varun Sharma, Thomas Krausgruber, Celine Sin, Jörg Menche, Eleni M. Tomazou, and Christoph Bock. Multimodal learning enables chat-based exploration of single-cell data. *Nature Biotechnology*, 2025. doi: 10.1038/s41587-025-02857-9.

Lucas Seninge, Ioannis Anastopoulos, Hongxu Ding, and Joshua Stuart. VEGA is an interpretable generative model for inferring biological network activity in single-cell transcriptomics. *Nature Communications*, 12(1):5684, 2021. doi: 10.1038/s41467-021-26017-0.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *International Conference on Learning Representations*, volume 2025, pp. 34635–34667, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/558d48c1f08675daa636e09bfe94a89e-Paper-Conference.pdf.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville Sanjana, and David Knowles. Faithful heteroscedastic regression with neural networks. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5593–5613. PMLR, 2023. URL https://proceedings.mlr.press/v206/stirn23a.html.

Brett Volmert, Artem Kiselev, Aniwat Juhong, Fei Wang, Ashlin Riggs, Aleksandra Kostina, Colin O'Hern, Priyadharshni Muniyandi, Aaron Wasserman, Amanda Huang, Yonatan Lewis-Israeli, Vishal Panda, Sudin Bhattacharya, Adam Lauver, Sangbum Park, Zhen Qiu, Chao Zhou, and Aitor Aguirre. A patterned human primitive heart organoid model generated by pluripotent stem cell self-organization. *Nature Communications*, 14(1):8245, 2023. doi: 10.1038/s41467-023-43999-1.

Vijaya Krishna Yalavarthi, Kiran Madhusudhanan, Randolf Scholz, Nourhan Ahmed, Johannes Burchert, Shayan Jawed, Stefan Born, and Lars Schmidt-Thieme. GraFITi: Graphs for forecasting irregularly sampled time series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16255–16263, 2024. doi: 10.1609/aaai.v38i15.29560.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer. In *International Conference on Learning Representations*, volume 2025, pp. 144–164, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/00b67df24009747e8bbed4c2c6f9c825-Paper-Conference.pdf.

Grace Hui Ting Yeo, Sachit D. Saksena, and David K. Gifford. Generative modeling of single-cell time series with PRESCIENT enables prediction of cell trajectories with interventions. *Nature Communications*, 12(1):3222, 2021. doi: 10.1038/s41467-021-23518-w.

Suyuan Zhao, Jiahuan Zhang, Yushuai Wu, Yizhen Luo, and Zaiqing Nie. LangCell: Language-cell pre-training for cell identity understanding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61159–61185. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24u.html.

A GENE DESCRIPTION SOURCES AND EXAMPLE INPUTS

Gene descriptions combine complementary information from three sources: NCBI Gene summaries, reviewed UniProt protein descriptions and Gene Ontology annotations. Table A1 illustrates these inputs for *MYBPC3* using excerpts from the cached records. Available source embeddings are averaged before the learned text projections; missing-source handling is described in Appendix B.

Table A1: Example gene-description inputs for *MYBPC3*.

Source	Example input text
NCBI Gene: 4607	"MYBPC3 encodes the cardiac isoform of myosin-binding protein C. … MYBPC3 is expressed exclusively in heart muscle and is a key regulator of cardiac contraction."
UniProt: Q14896	"In vitro it binds MHC, F-actin and native thin filaments, and modifies the activity of actin-activated myosin ATPase."
Gene Ontology	Biological process: cardiac muscle contraction. Molecular function: structural constituent of muscle. Cellular component: striated muscle myosin thick filament.

Excerpts preserve the cached input wording; the ellipsis marks omitted text. The NCBI summary is from the cached GenePT distribution, and UniProt and GO excerpts are from frozen annotation records. MHC denotes myosin heavy chain.

B ARCHITECTURE AND IMPLEMENTATION DETAILS

This appendix specifies the architecture in Section 3 and Figure 2.

B.1 INPUT NORMALIZATION AND GENE FEATURES

Both branches share the task-specific expression input $\mathbf{g}_{t_k}^{\mathrm{in}}$. We apply gene-wise normalization to obtain the shared expression context:

$$g_{t_k,j}^{\mathrm{ctx}} = \widetilde{g}_{t_k,j}^{\mathrm{in}} = \frac{g_{t_k,j}^{\mathrm{in}} - \bar{g}_j^{\mathrm{train}}}{s_j},$$

where the gene-wise means $\bar{g}_j^{\mathrm{train}}$ and scales s_j are estimated from permitted training profiles only. Gene identifiers align expression, text and pathway-membership records. Learned identity embeddings distinguish pathways and auxiliary components.

For each training setting, the time scale s_t is the median positive gap between consecutive observations within recorded groups during the permitted training period. It remains fixed throughout training and forecasting.

We average the available source vectors to obtain $\mathbf{t}_j$, setting it to zero when all three sources are missing. The Gene Transformer applies its own learned linear projection. A separate adapter constructs $\mathbf{f}_j \in \mathbb{R}^d$ from text and numeric contributions. The text contribution subtracts the adapter's output at zero input from its output at $\mathbf{t}_j$, so missing text contributes zero. The numeric adapter uses $\operatorname{asinh}(\bar{g}_j^{\mathrm{train}})$, $\log s_j$ and the three binary source-availability indicators. Both adapters are optimized jointly with the forecasting model, while the source vectors remain fixed. Appendix A illustrates the text sources.

B.2 PATHWAY-AWARE GENE ENCODING

In Equation 1, the expression mapping is

$$\phi_{\mathrm{expr}}(x) = \mathbf{w}_{\mathrm{expr}} \operatorname{asinh}(x) + \mathbf{b}_{\mathrm{expr}},$$

where x is a standardized expression scalar, $\mathbf{w}_{\mathrm{expr}}$ and $\mathbf{b}_{\mathrm{expr}}$ are learned vectors, and the inverse hyperbolic sine preserves sign while compressing large magnitudes. The learned matrix $\mathbf{W}_{\mathrm{txt}}$ projects

the text vector into the same space. We add the expression and text representations and apply non-affine LayerNorm to form the gene token.

Standard self-attention projects queries, keys and values from the expression–text tokens, with parameters shared across pathways and times. Both sublayers use residual connections, with an additional non-affine LayerNorm before the feed-forward sublayer. All measured pathway members are retained; padding is excluded from attention keys and mean pooling. A shared linear readout produces $d_{\mathcal{P}}$ features per pathway and is zero-initialized in the paired encoder-deletion configuration. The Gene Transformer replaces the original masked expression-to-pathway encoder and its additive bypass; the reconstruction dictionary remains trainable.

B.3 History encoding and time conditioning

Let $\mathbf{x}_{k,i}$ denote a pathway embedding $\mathbf{x}_{k,i}^{(\mathcal{P})}$ or a dimension-aligned auxiliary token. We construct the history token by concatenating this representation with its adjacent-time slope $\dot{\mathbf{x}}_{k,i}$, running mean $\bar{\mathbf{x}}_{k,i}$ and normalized sampling-time features $\boldsymbol{\tau}_k$:

$$\mathbf{e}_{k,i} = \phi_{\text{in}}([\mathbf{x}_{k,i}; \dot{\mathbf{x}}_{k,i}; \bar{\mathbf{x}}_{k,i}; \boldsymbol{\tau}_k]) + \mathbf{e}_i^{\text{ID}}. \tag{10}$$

Here, ϕ_{in} maps the concatenated features to d dimensions, and $\mathbf{e}_i^{\text{ID}}$ is a learned pathway or auxiliary-component identity embedding.

Both the slope and running mean use observations through t_k only:

$$\dot{\mathbf{x}}_{k,i} = \frac{\mathbf{x}_{k,i} - \mathbf{x}_{k-1,i}}{(t_k - t_{k-1})/s_t}, \qquad \bar{\mathbf{x}}_{k,i} = \frac{1}{k}\sum_{r=1}^{k} \mathbf{x}_{r,i},$$

The first observation has zero slope. The time features are $\boldsymbol{\tau}_k = [\Delta t_k/s_t, (t_k - t_{\text{obs}})/s_t, t_k/s_t]$. Together, the embedding, slope, mean and time features form the vector $\mathbf{v}$. The input mapping applies an element-wise transform before the MLP:

$$\phi_{\text{in}}(\mathbf{v}) = \text{MLP}_{\text{in}}(\text{asinh}(\mathbf{v}))\,.$$

Here, MLP_{in} uses GELU activations. Dictionary-projected auxiliary states are zero-padded to the pathway embedding width before token construction. No centered-deviation feature is included.

The temporal and cross-pathway Transformers each comprise N layers. Temporal attention is bidirectional within the supplied observed prefix. Only valid observations are processed: padding is masked, and future target profiles are excluded.

Forecast-time information enters through the embedding in Equation 3:

$$\boldsymbol{\eta}(t_{\text{obs}}, \widetilde{\Delta t}) = \phi_{\text{time}}\left(\text{asinh}\left[\frac{\widetilde{\Delta t}}{1+\widetilde{\Delta t}}; \log(1+\widetilde{\Delta t}); \frac{t_{\text{obs}}}{s_t}\right]\right),$$

The mapping ϕ_{time} is an MLP with GELU. Its output is added to the final temporal representation before cross-pathway attention. Evaluating the same module at zero and at the requested horizon gives $\mathbf{h}_i(0)$ and $\mathbf{h}_i(\widetilde{\Delta t})$, respectively.

B.4 Differential attention

For a single head with input token matrix $\mathbf{X}$, differential self-attention (Ye et al., 2025) is

$$\text{DiffAttn}(\mathbf{X}) = \left[\text{softmax}\left(\frac{\mathbf{Q}_1\mathbf{K}_1^\top}{\sqrt{d_h}}\right) - \lambda\,\text{softmax}\left(\frac{\mathbf{Q}_2\mathbf{K}_2^\top}{\sqrt{d_h}}\right)\right]\mathbf{V}. \tag{11}$$

The two query–key pairs $(\mathbf{Q}_1, \mathbf{K}_1)$ and $(\mathbf{Q}_2, \mathbf{K}_2)$ and shared values $\mathbf{V}$ are learned projections of $\mathbf{X}$. The query–key dimension per head is d_h, and λ controls the subtraction. Softmax is applied row-wise.

The subtraction coefficient is

$$\lambda = \exp(\boldsymbol{\lambda}_{q1}^\top \boldsymbol{\lambda}_{k1}) - \exp(\boldsymbol{\lambda}_{q2}^\top \boldsymbol{\lambda}_{k2}) + \lambda_{\text{init}}, \qquad \lambda_{\text{init}} = 0.8 - 0.6\exp(-0.3\ell),$$

where $\ell \in \{0, \ldots, N-1\}$ is the zero-based layer index within each stack. The four learned vectors are independently initialized from $\mathcal{N}(0, 0.1^2)$ and shared across heads within a layer. Each head's difference output is RMS-normalized with $\epsilon = 10^{-5}$ and a learned scale, then multiplied by $1-\lambda_{\text{init}}$. The head outputs are concatenated and linearly projected.

The Transformer uses pre-LayerNorm and residual connections around attention and the GELU feed-forward sublayer. Differential attention is used for temporal and cross-pathway modeling; gene-level encoding uses standard self-attention. We adopt the attention operator without scDFM's perturbation-conditioned flow-matching objective.

B.5 State reconstruction and normalization

The reconstruction dictionary has $N_z = KN_{\mathcal{P}} + N_{\text{aux}}$ rows and $N_{\mathcal{G}}$ columns, with K coordinates per pathway state and one scalar state per auxiliary component. Pathway rows combine learned base weights with a bounded adjustment from $\mathbf{f}_j$, retain member connections only, and are normalized to the square root of their member count. Dictionary parameters are separate from the Gene Transformer.

We initialize pathway rows from the leading within-pathway singular vectors of standardized training profiles and auxiliary rows from the residual expression subspace. All dictionary weights receive both forecasting and reconstruction gradients. The joint projection in Equation 4 uses $\lambda_{\text{proj}} = 0.1$ and $\Lambda_{aa} = \max\{(\mathbf{W}^{\text{base}}(\mathbf{W}^{\text{base}})^\top)_{aa}, 1\}$. Jointly inferred states can depend on nonmember genes despite the membership constraints on decoder connections.

The state branch summarizes the latest state and its recent slope. For $T \geq 2$,

$$\dot{\mathbf{z}}_i^{\text{recent}} = \frac{\mathbf{z}_{t_T,i} - \mathbf{z}_{t_{T-1},i}}{(t_T - t_{T-1})/s_t}.$$

A one-observation prefix has zero slope. To normalize each state component a, we compute its RMS over the observed prefix:

$$s_{ia}^z = \max\left\{ \sqrt{T^{-1}\sum_{k=1}^{T} z_{t_k,ia}^2},\ 0.01 \right\}.$$

Element-wise division by $\mathbf{s}_i^z$, followed by the inverse hyperbolic sine, gives the state summary

$$\mathbf{u}_i = \left[\operatorname{asinh}(\mathbf{z}_i^0/\mathbf{s}_i^z); \operatorname{asinh}(\dot{\mathbf{z}}_i^{\text{recent}}/\mathbf{s}_i^z)\right] \in \mathbb{R}^{2K}. \tag{12}$$

Auxiliary components follow the same construction, with each scalar state and slope zero-padded to width K for the shared heads. This summary conditions the forecast heads, whose amplitudes and velocities are mapped back to the state scale as described below. The summary uses the recent slope and excludes the full-span slope.

B.6 Forecast heads and expression decoding

Trajectory parameters. The trajectory heads receive $\mathbf{c}_i(0) \in \mathbb{R}^{d+2K}$. Level amplitudes and initial and sustained trend velocities are scaled by the corresponding observed-prefix RMS in $\mathbf{s}_i^z$. The offset δv_{ia} equals the initial velocity minus the sustained velocity v_{ia}. A sigmoid produces α_i, shared across the pathway's K state components, and softplus plus 10^{-4} ensures positive time constants.

Dynamic pathway decoding. The history context $\mathbf{h}_i(\widetilde{\Delta t})$ and gene features $\mathbf{f}_j$ are separately mapped to bounded vectors of matching dimension. Their mean element-wise product, scaled by 0.15, adjusts each pathway–gene weight across all K dictionary rows. We add the adjustment to $\mathbf{W}_i^{\text{base}}$, reapply the membership mask and normalize rows to the square root of their member counts. The resulting $\mathbf{W}_i(\widetilde{\Delta t})$ decodes $\Delta\mathbf{z}_i$ through Equation 6. Its zero-horizon counterpart $\mathbf{W}_i(0)$ provides the regularization reference in Appendix C. Level–trend parameters use $\mathbf{c}_i(0)$ and are shared across horizons for a given history.

Auxiliary and gene-level contributions. The dense auxiliary block of the base dictionary maps auxiliary-state increments to expression changes:

$$\Delta\widehat{\mathbf{g}}^{\mathrm{aux}} = \mathbf{s} \odot (\mathbf{W}^{\mathrm{aux}})^{\top} \Delta\mathbf{z}^{\mathrm{aux}}. \tag{13}$$

For the gene-level correction, a shared MLP combines $\mathbf{f}_j$ with the latest three available expression values, recent and full-span slopes, the observed mean, reconstruction residuals and observation-time features. It predicts a signed amplitude b_j and a positive time constant τ_j^{gene}, yielding

$$\Delta\widehat{g}_j^{\mathrm{gene}} = s_j b_j \kappa(\widetilde{\Delta t}; \tau_j^{\mathrm{gene}}) = s_j b_j \left[1 - \exp(-\widetilde{\Delta t}/\tau_j^{\mathrm{gene}})\right]. \tag{14}$$

B.7 Training configuration

We train PathCast using AdamW with a learning rate of 10^{-3}, weight decay of 0.01, and gradient-norm clipping at 1.0. Training queries are grouped by trajectory into batches of up to 64 queries. Normalization statistics are computed from the permitted training observations and remain fixed during forecasting. Experiments use random seeds 17, 42, and 73, and results are averaged across the three runs.

C Training objective and regularization

The objective in Equation 8 combines prediction, trajectory and reconstruction losses with regularization of output contributions and decoder adjustments.

Prediction, dynamics and reconstruction losses. For training query n, let $\boldsymbol{\varepsilon}_n = \widehat{\mathbf{g}}_n - \mathbf{g}_n$. Query weights w_n balance recorded groups and observed prefixes and sum to one. For queries sharing an observed prefix, $\mathcal{A}$ contains adjacent target pairs, while $\mathcal{S}$ contains the first–last pairs for prefixes with at least three targets. We denote the elapsed time between paired targets by $\Delta t_{nn'}$. The set $\mathcal{H}$ indexes observed history profiles, with standardized inputs $\mathbf{g}_{n,k}^{\mathrm{ctx}}$ and latent states $\mathbf{z}_{n,k}$. The loss components are

$$\begin{aligned}
\mathcal{L}_{\mathrm{pred}} &= \frac{1}{N_{\mathcal{G}} c_{\mathrm{pred}}} \sum_n w_n \|\boldsymbol{\varepsilon}_n\|_2^2, \\
\mathcal{L}_{\mathrm{dyn}} &= \lambda_{\mathrm{rate}} \mathcal{L}_{\mathrm{rate}} + \lambda_{\mathrm{growth}} \mathcal{L}_{\mathrm{growth}}, \\
\mathcal{L}_{\mathrm{rate}} &= \frac{1}{|\mathcal{A}| N_{\mathcal{G}} c_{\mathrm{rate}}} \sum_{(n,n')\in\mathcal{A}} \left\| \frac{\boldsymbol{\varepsilon}_{n'} - \boldsymbol{\varepsilon}_n}{\Delta t_{nn'}} \right\|_2^2, \\
\mathcal{L}_{\mathrm{growth}} &= \frac{1}{|\mathcal{S}| N_{\mathcal{G}} c_{\mathrm{growth}}} \sum_{(n,n')\in\mathcal{S}} \|\boldsymbol{\varepsilon}_{n'} - \boldsymbol{\varepsilon}_n\|_2^2, \\
\mathcal{L}_{\mathrm{recon}} &= \frac{1}{|\mathcal{H}| N_{\mathcal{G}} c_{\mathrm{recon}}} \sum_{(n,k)\in\mathcal{H}} \left\| \mathbf{s} \odot \left[(\mathbf{W}^{\mathrm{base}})^{\top} \mathbf{z}_{n,k} - \mathbf{g}_{n,k}^{\mathrm{ctx}} \right] \right\|_2^2.
\end{aligned} \tag{15}$$

Rate and growth losses supervise local and overall trajectory changes, respectively. Reconstruction is measured on the input-expression scale.

Loss normalization and weights. The positive normalizers c_{pred}, c_{rate}, c_{growth} and c_{recon} are training-set mean-square scales. They correspond, respectively, to target increments relative to $\mathbf{g}^{\mathrm{base}}$, adjacent rates, first–last changes and centered input expression. The weights λ_{rate}, $\lambda_{\mathrm{growth}}$ and λ_{recon} control the contributions of the corresponding loss terms. All pairs and normalization scales are constructed from permitted training observations.

Output and decoder regularization. For the current hybrid configuration, $\mathcal{R}$ constrains output amplitudes and decoder-weight adjustments as follows:

$$\mathcal{R} = \lambda_{\mathrm{conn}} \mathcal{R}_{\mathrm{conn}} + \frac{\lambda_{\mathrm{res}} \mathcal{R}_{\mathrm{res}} + \lambda_{\mathrm{trend}} \mathcal{R}_{\mathrm{trend}} + \lambda_{\mathrm{trans}} \mathcal{R}_{\mathrm{trans}}}{c_{\mathrm{pred}}}.$$

Here, λ_{conn}, λ_{res}, λ_{trend} and λ_{trans} are regularization weights. $\mathcal{R}_{\text{res}}$ sums the mean-square auxiliary and gene-level expression corrections. The term $\mathcal{R}_{\text{trend}}$ sums the mean-square expression contributions of individual pathway trends, while $\mathcal{R}_{\text{trans}}$ penalizes level contributions. Both pathway terms use the increment decoder in Equation 6. The connection penalty $\mathcal{R}_{\text{conn}}$ averages squared decoder adjustments relative to the base and origin decoders, normalized by pathway member counts. The origin-to-target adjustment is additionally divided by $1 + \widetilde{\Delta t}^{\,2}$.

The objective includes adjacent-rate and first–last-change supervision, with no additional origin-rate or pathway-average prediction loss. The first–last-change loss remains active when the full-span slope is excluded from state conditioning.

D Predictive uncertainty and interval calibration

D.1 Residual-variance estimation

A shared variance head F_{var} combines gene features, expression-history statistics, forecast increments and time features, including the forecast horizon. An MLP with a positive output parameterization produces the residual variance $\sigma_j^2(t)$. For each permitted training target $g_{t,j}$, we minimize the Gaussian negative log-likelihood, up to an additive constant:

$$\ell_{\text{var},j}(t) = \frac{1}{2}\left[\log \sigma_j^2(t) + \frac{\{g_{t,j} - \widehat{g}_{t,j}\}^2}{\sigma_j^2(t)}\right].$$

The expression prediction $\widehat{g}_{t,j}$ and the head's inputs are detached from the forecasting computation graph, so this objective updates only the variance head (Stirn et al., 2023). The forecasting model is optimized with Equation 8, and both objectives can be fitted concurrently. The variance head estimates residual scale without changing the expression forecast; variation across independently fitted models is handled by the ensemble extension below.

D.2 Calibration from historical residuals

We calibrate the intervals in Equation 9 using historical forecast residuals pooled across recorded groups and genes. Let $\mathcal{C}$ contain the historical target–gene pairs, with the group index omitted for clarity. For nominal coverage p, the calibration multiplier is the empirical p-quantile of the absolute standardized residuals:

$$q_p = \text{Quantile}_p\left(\left\{\frac{|g_{t,j} - \widehat{g}_{t,j}|}{\max\{\sigma_j(t), 10^{-8}\}} : (t,j) \in \mathcal{C}\right\}\right).$$

We use the empirical `higher` quantile, which selects the next observed order statistic. This recalibration adjusts interval widths while leaving expression predictions unchanged (Kuleshov et al., 2018). Residuals must come from the predictor being evaluated, so ensemble multipliers cannot serve as calibration results for a single model. Pooling residuals also provides no separate coverage guarantee for individual genes.

D.3 Ensemble uncertainty estimation

For M_{ens} independently fitted models, let $\widehat{g}_{t,j}^{(m)}$ and $\sigma_{m,j}^2(t)$ denote model m's expression prediction and residual variance. The ensemble prediction averages the individual forecasts, $\widehat{g}_{t,j}^{\text{ens}} = M_{\text{ens}}^{-1}\sum_{m=1}^{M_{\text{ens}}} \widehat{g}_{t,j}^{(m)}$, and its variance combines the average residual variance with variation among model predictions (Lakshminarayanan et al., 2017):

$$\sigma_{*,j}^2(t) = \frac{1}{M_{\text{ens}}}\sum_{m=1}^{M_{\text{ens}}}\left[\sigma_{m,j}^2(t) + \{\widehat{g}_{t,j}^{(m)} - \widehat{g}_{t,j}^{\text{ens}}\}^2\right].$$

Calibration follows the same rule, replacing the single-model prediction and standard deviation with $\widehat{g}_{t,j}^{\text{ens}}$ and $\sigma_{*,j}(t)$. The EB-hCO analysis uses three models.

E EVALUATION SETTINGS AND CROSS-STUDY TRANSFER

Table E1 summarizes the observed histories and forecast windows used in each evaluation. T3 excludes EB-hCO from source pretraining and adapts the pretrained model using either of two EB-hCO prefixes, without target-time reference profiles. H1-Hep provides an additional evaluation on hepatic differentiation; its Top-50 genes are selected by expression variance across the observed history.

Table E1: Observed histories and forecast windows across evaluation settings.

Task	Held-out axis	Evaluation units	Observed days	Forecast days
T1	Future time	EB-hCO	D0, D1, D3, D5, D7, D9, D11	D13, D15, D17, D19
		Cardioids CHIR4	D0, D1.5, D2.5, D3.5, D4.5, D5.5	D6.5, D7.5, D10
		Cardioids CHIR8	D0, D1.5, D2.5, D3.5, D4.5, D5.5	D6.5, D7.5, D10
T2	Condition effect	cVO Unvasc.↔Vasc.	D2, D5, D8	D10, D12, D16
		Cardioids CHIR4↔CHIR8	D0, D1.5, D2.5, D3.5, D4.5, D5.5	D6.5, D7.5, D10
T3	Independent study	EB-hCO (through D11)	D0, D1, D3, D5, D7, D9, D11	D13, D15, D17, D19
		EB-hCO (through D7)	D0, D1, D3, D5, D7	D13, D15, D17, D19
Additional	Future time	H1-Hep	D0, D1, D2, D3, D5, D7, D9	D11, D13

E.1 TRANSFER PERFORMANCE ACROSS FORECAST TIMES

We examine how cross-study transfer affects accuracy at individual forecast times by comparing frozen transfer, full-parameter adaptation and prediction-head adaptation with training from scratch. As shown in Figure E1, both adaptation strategies reduce MSE at every forecast day for both observed prefixes. Frozen transfer also improves most forecasts, with the exception of D13 when the history extends through D11. With the shorter D7 history, all three transfer settings reduce error at every forecast day.

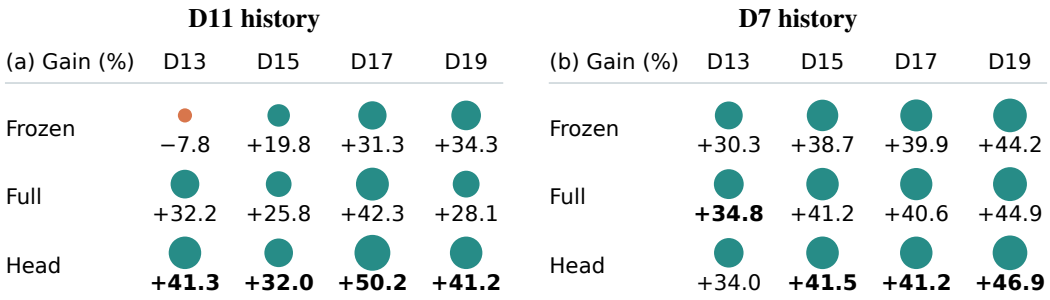

Figure E1: Cross-study transfer gains on EB-hCO with histories through (a) D11 and (b) D7. Values report percentage MSE reductions relative to training from scratch. Circle area indicates magnitude; teal denotes lower error and orange higher error. Bold marks the largest gain at each forecast time.

F PATHWAY CONTRIBUTIONS AND STATE DYNAMICS

Stability of gene contributions. We assess whether the leading gene contributions in Figure 5 are consistent across seeds. For each pathway, four of the five leading D19 contributors remain in the top five across all three seeds. COL3A1 in Myogenesis and ACTN2 in Striated muscle contraction

appear in the top five in two seeds. The output-removal analysis uses five fixed control gene sets matched by annotation resource, size, coverage and gene sharing. Each control's effect is averaged across seeds; diamonds summarize the median and min–max range across controls.

Pathway-state readouts and forecast dependence. Figure F1 compares internal state readouts with measured pathway RNA scores for five pathways selected after inspecting sixteen candidates. We compute each score by averaging gene-standardized member expression, then standardize it using D0–D11 score statistics. Linear state readouts are fitted on D0–D11 and anchored at the observed D11 score. Bars average recorded groups and seeds; error bars show ± 1 SD across three seeds after group averaging. To assess forecast dependence on each pathway, we set its state increment to zero while retaining its target-time decoder and all other outputs. MSE changes are computed across all 2,000 genes within each seed and forecast day, then averaged across seeds. These readouts are expression proxies, and both removal analyses characterize the fitted model rather than establish biological causation.

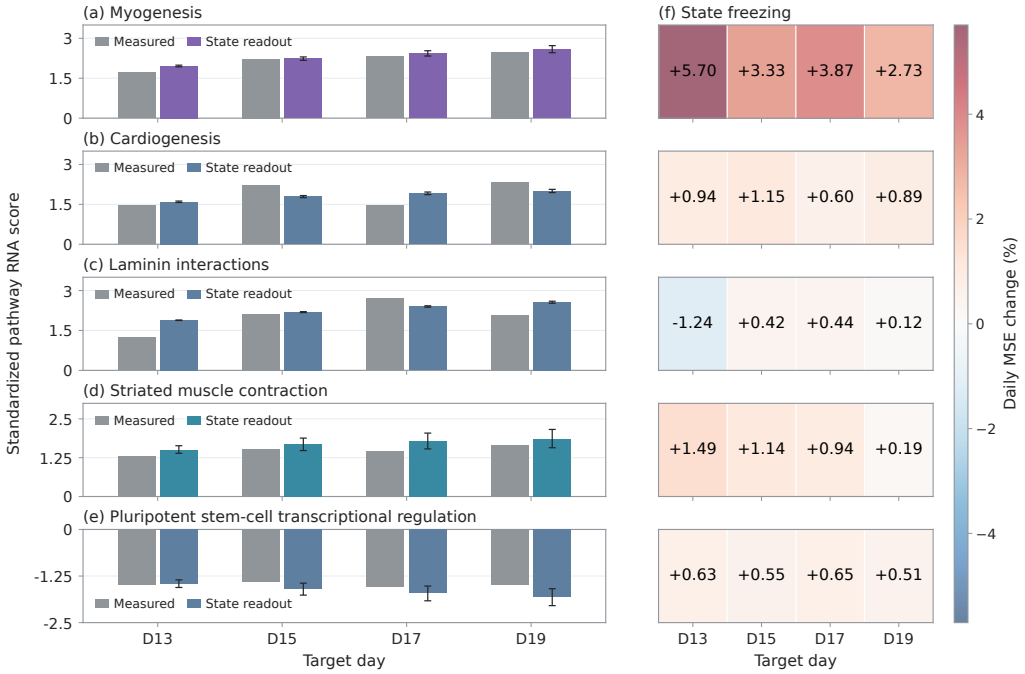

Figure F1: Pathway-state readouts and forecast dependence on EB-hCO. (a–e) Measured pathway RNA scores (gray) and state readouts (colored). (f) Percentage change in MSE after setting each pathway's state increment to zero; positive values indicate increased error.